MULTIMODAL ATTENTION NETWORKS FOR 5G DEEPSENSE

by

ANDREW EL KOMMOS
M.S. University of Central Florida, 2024

A thesis submitted in partial fulfilment of the requirements
for the degree of Master of Engineering
in the Department of Electrical Engineering
in the College of CECS Engineering
at the University of Central Florida

Spring Term
2024

ABSTRACT

The abstract page should be an essay-style summary of the purposes, methodology, findings or conclusions. It should not contain tables or figures of any kind. It is double-spaced. The heading ABSTRACT should be centered, without punctuation, at the top margin. If more than one page is necessary, continue on the following page. Do not repeat the heading or use the word continued.

The past X years have seen an large advancement in AI systems capable of solving complex tasks.

Beam Management for millimeter wave and sub-terahertz communication systems remains a challenging task in dense urban environments. The approach of using real world sensing has attracted interested in the field of communications from both Academia and industry, which has spawned The "Multi-Modal Beam Prediction Challenge 2022: Towards Generalization" competition. This competition aims to offer a platform for investigating the viability of aligned multi-modal sensing in aiding in generalized beam management for real-world future communication systems. This research aims to apply deep learning principles in self-supervised learning and attention based networks in the aims to produce a multi-modal deep learning framework for effective generalization of the 5G/6G Beam Management problem outlined in the 2022 challenge to a multitude of sensing scenarios.[**?**]

I dedicate this thesis to my parents, and siblings for supporting me to continue through my education...etc

ACKNOWLEDGMENTS

The acknowledgments page is optional. If you choose to use it, it should appear after the Abstract, but before the Table of Contents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Chapter and major headings should be typed in all caps. Note that Chapter titles should be formatted and positioned exactly the same as frontmatter and other major headings. However, chapters with subtitles may be stacked, single-spaced, rather than appear on one line. The Introduction presents an overview of the thesis or dissertation material to be discussed. For sample theses and dissertations, including sample Introductions from your discipline, visit the University Writing Center's Graduate Gateway, located at http://www.uwc.ucf.edu. Please be aware that UWC links are for content samples only, not format samples.

Background

Global communication networks have matured quite significanly since the first 2G mobile radio networks were deployed in the 1990s. TO support increasing network requirements for an increase user base, and throughput the communication systems will need to continue to mature. It is projected by 2030, global mobile traffic will be 670-times of the traffic in 2010, mainly due to machine-to-machine (M2M) communications. [**?**][**?**] To support this unprecedented exponential growth networks beyond 5G will need developed.

The "Deepsense 6G Dataset" is a large-scale dataset based on real-world measurements of co-existing multi-modal sensing and communication data. It compromises co-existing and synchronized multi-modal sensing and communication and is organized in a collection [**?**]

Problem Statement

Multimodal Networks

The field of multimodal machine learning aims to develop models that can process and integrate data from several modalities. A unimodal can be represented by

$$D_u = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

In this equation, each pair (x_i, y_i) represents the input and corresponding label for the i-th data point, with i ranging from 1 to n. [**?**]

Whereas a multimodal network would include data from multiple sensing modalities such as audio, image, radar, or text. This information can be shared or complementary data and can be represented as

$$D_m = \left\{(x_1^1, \ldots, x_1^k, y_1), \ldots, (x_n^1, \ldots, x_n^k, y_n)\right\}$$

where k is the number of input modalities, and the corresponding label for the i-th data point ranging from 1 to n. Which would represent paired multimodal data. [**?**]

Objectives

CHAPTER 2: LITERATURE REVIEW

Communication Systems

As modern communication systems move towards the deployment of 6G networks, the need for efficient and reliable communication systems has become increasingly important. The use of machine learning techniques in communication systems has been shown to improve the performance of wireless communication systems by optimizing the use of resources and reducing the latency of data transmission.

To address challenges in 6G communication systems novel techniques such as reconfigurable intelligence surfaces (RISs), integrated sensing and communication, mmWave and THz networks, and other methods have been proposed. [**?**] Although these methods demonstrate satisfactory performance towards 6G requirements, the requirements for network management have increased in complexity and scale. The future 6G networks have leveraged machine learning (ML) as promising solution to network management techniques which optimize the use of resources and reduce the latency of data transmission. [**?**] To address the challenges in communication systems complexity there have been several studies that leverage techniques such as reinforcement learning [**?**], map assisted localization ray tracing techniques [**?**], deep network assisted channel state information (CSI) prediction [**?**]. These studies have demonstrated the potential of machine learning techniques in improving the performance of wireless communication systems.

Recently, large language models (LLM) techniques have demonstrated the potential to improve through the use of massivily large models that demonstrate the ability to reason and comprehend complex tasks across various domains. The emergence of these models have shown to be effective in addressing a number of different challenges in the domains of health care, law, finance,

education, and other technical fields.

From early in the development of language models, natural language research such as, GPT-2 showed that LLM technologies have emergent capabilities and are unsupervised multitask learners. [**?**]. Since this seminal paper by Radform, et. al, the field of large language models has matured significantly, and new research has emerge on how to apply the capabilities of these models across various domains and different ways. The technique of (prompt engineering) is a technique that gained prominence and served as a way to improve the performance of a model by crafting input prompts. By altering the phrasing or structure of prompts, the LLM can be fine-tuned to align the behavior with a desired task. [**?**] The concept of In-Context learning was explored in [**?**] and demonstrated how LLMs can adapt to new tasks by utilizing contextual clues in the input prompt. To build on the capabilities of utilizing LLMs without the need for fine-tuning, the concept of *(*Chain of Thought reasoning) CoT was introduced in [**?**] and demonstrates how LLMs can guided through a sequence of intermeidate reasoning steps to solve complex problems. The CoT reasoning technique demonstrated substantial improvements in solving mathematical and logical reasoning tasks by breaking down the problem into smaller steps.

Prior Deepsense Methods

The Deepsense 6G dataset is a real-world multi-modal dataset that provides researchers a comprehensive dataset for multi-modal sensing and communication data. The dataset provides the world's first large-scale real-world sensing and communication reprository of over a million data points, with over 30 different scenarios that target multiple applications. The team that provided the DeepSense 6G dataset aims to provide a platform that encourages the development of machine learning solutions for applications in communication systems, through a variety of multi-modal sensor technologies. There are various tasks that are supported by the dataset as well, ranging

from *Sensing Aided Beam Prediction* to *Future Blockage Prediction*, with each task supported by a diverse set of sensing modalities. The dataset provides multiple input modalities such as GPS, Camera, Radar, and LiDAR sensing modalities aimed to enhance the breadth of the dataset for various applications. In addition to the various data points, test benches, and scenarios, the Deepsense team provides a wide array of papers, code bases, and techniques for solving problems relevant to the dataset. The inception of the DeepSense 6G dataset has marked a significant milestone in fostering research within the field of communications, particularly in scenarios requiring multi-modal sensor integration. Before the introduction of datasets like DeepSense 6G, researchers encountered substantial challenges, predominantly due to the lack of access to large-scale, realistic datasets that reflect the complexity and variety of real-world environments. [**?**]

Traditional methods for beam management often rely on exhaustive beam search techniques, which often have drawbacks in high overhead and an intractible search space. Leveraging a large-scale dataset Machine Learning frameworks are viable alternative to traditional datascience techniques as outlined in the paper, *Computer Vision Aided Beam Tracking in a Real-World Millimeter Wave Deployment*. In this work the authors propose to utilize temporal visual sensing information to predict the optimal beam, as defined as the highest beamforming gain through the use of an Encoder-Decoder network. The paper presents an innovative machine learning (ML) model for beam tracking optimization in mmWave communication systems, in which a base station is equipped with an antenna array and an RGB camera to assist mobile users. The proposed model leverages visual sensing information alongside pre-defined beamforming codebooks to predict optimal future beams using an encoder-decoder architecture leveraging Recurrent Neural Networks (RNNs). The vision-aided approach demonstrates promise over the traditional methods by leveraging existing feature extraction networks to enhance beamforming gain and communication performance. [**?**] The paper underscores the effectiveness of machine learning-based and vision-aided beam tracking in mmWave communications. This approach shows that it is possible to achieve a high level of

accuracy for narrowing the beam search to top-5 accuracy with a precision of 99.37 percent for the next beam by observing previous time-steps on a single scenario.

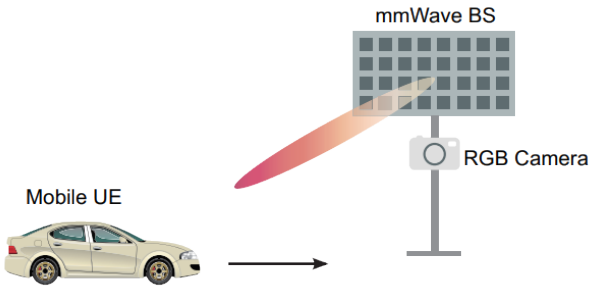

Figure 2.1: The considered system model leveraged to design a computer vision aided Beam Tracking system [**?**].

The paper's focus on next beam prediction presents limited scope, and does not take into account the some of the deeper challenges often faced in wireless communication systems by limiting it's evaluation to a single scenario. In turn the paper fails to consider the variability of real-world environments. The omission of additional sensing modalities such as radar, GPS, or other sensing inputs might limit the system's capabilites in Non-Line of Sight conditions (NLOS), where additional information might be crucial. Other methods delve into the usage of position-aided beam prediction aiming to evaluate the practicality of GPS guided predictions. In the paper *Position-aided Beam Prediction in the Real World: How Useful GPS Locations Actually Are?*, the authors propose three machine learning solutions for position-aided mmWave beam prediction. They evaluate multiple methods using a Lookup Table, K-Nearest Neighbors, and a fully connected Neural network against three different scenarios. The paper reveals that the Neural Network generally outperforms both the lookup table and K-Nearest Neighbors in top-1 predictions across various scenarios due to the ability of the Neural Network to generalize better and utilize more information from training samples. They do note that factors such as input noise and label noise degrades performance and necessitates alternate metrics such as power loss, but the paper outlines the prac-

tical implications of using position data for beam alignment. Additionally, the authors point out that utilizing GPS positions can provide significant savings for the overhead of beam search in a code book that contains 64 unique codes. [**?**]. There are several drawbacks to the use of GPS for beam alignment and some of these challenges include susceptibility to noisy inputs, latency issues, and environmental constraints, which lead the authors to suggest techniques that do not solely rely on GPS based methods.

Location-aware methods offer a practical approach but come with limitations such as latency issues, environmental blockages, and noisy inputs. Works by the DeepSense team explore the efficacy of using Radar Aided Beam Prediction in the paper *Radar Aided 6G Beam Prediction: Deep Learning Algorithms and Real-World Demonstration*. The team proposes a novel radar-aided deep learning framework to map radar samples to the optimal beam predictions. The framework leverages preprocessing of the radar samples into feature maps such as range, angle, and velocity maps using Fast Fourier Transforms (FFTs) and then employ a deep neural network subsequently to learn the mapping from the features to optimal beamforming predictions. The network is designed to be a relatively simple deep network comprosing of convolutional and fully-connected layers utilizing the rectified linear unit (ReLU) non-linear activation functions. The final layer of the network maps a hidden-layer to the designed beam code book size of 64 positions in which the objective function utilizes cross-entropy loss for predicting the optimal beam. The promise for utilizing relatively simple deep learning networks and demonstrates the capabilities of the approach showing how a the model can acheive around 90 percent for top-5 accuracy. This research underscores the potential for utilizing radar for inferring beam prediction with the radar modality. [**?**]

In addition to radar, vision, or position based sensing methods, other modalities like LiDAR (Light Detection and Ranging) can contibute to the diversification of environmental perception. *LiDAR Aided Future Beam Prediction in Real-World Millimeter Wave V2I Communications* explores the effectiveness of utilizing LiDAR in the Beam prediction and reduction tasks. The authors argue

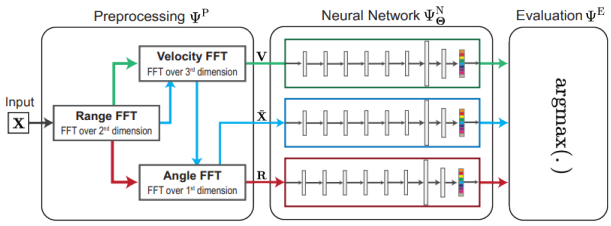

Figure 2.2: The considered architecture leveraged implemented in the radar aided Beam tracking paper. [**?**].

that vision-aided methods might fall short in low light conditions, and may cause privacy concerns when it comes to positional information of the end user. The paper investigates utilizing LiDAR for mmWave beam prediction and tracking tasks, which in turn shows promising results in predicting optimal beam at 95 percent for top 5 accuracy.

Large Language Models in Telecommunications

Previous studies have made significant contributions to addressing the Deepesense challenge problem sets, these methods primarily focused using deep learning techniques for millimeter-wave based communication systems. Researchers have proposed various machine learning-based approaches to address the challenges of wireless communication systems, however a these techniques are often limited by their reliance on dataset specific features. The reliance on pre-trained object detection models, such as YOLOv3, or Resnet50, may limit the network's ability to generalize to new, unseen data or "zero-shot" tasks in which the model has not been trained on similar data. In contrast, large language models have been shown to demonstrate emergent properties that allow them to generalize to new tasks and data. As demonstrated in [**?**] language models such as GPT-2 have been shown to perform well on a wide range of tasks, including text generation, summarization, and question answering. These foundational models pose a significant advantage over

traditional deep learning models due to their ability to solve tasks in zero-shot, flexibility and capability to multitask, ability to transfer domains or unseen data, and their ability to generalize across different task sets. These generalization capabilities are particularly important in the context of wireless communication systems, where the environment is constantly changing and evolving. In this section, we will review the literature on large language models and their applications in the field of wireless communication systems.

Multimodal networks

Transformer Architectures

Vision Language Models

CHAPTER 3: METHODOLOGY

Data Preprocessing

The Deepsense 5G dataset compromises of data points aligned modalities using video, radar, lidar, gps, and radar.

Architecture

Training Setup

Chapter Three, also called “Methodology,” “Research Methods,” or “Research Design and Methodology,” generally presents an overview of the methods used for researching the subject.

Numbering Subheadings

All appearances of those numbered headings and subheadings, including the Table of Contents and the bookmarks, should feature exactly the same language, numbering and formatting.

CHAPTER 4: Results

Chapter Four, also called “Results” or “Data Analysis,” usually provides detailed findings of the research. This chapter is where tables and figures most often appear, so make sure formatting is consistent.

Sample Table

The following sample table is an example of acceptable table formatting. Descriptive titles appear above tables and may appear either on one line or stacked and single-spaced. The table itself may also be single-spaced as necessary. If at all possible, try to keep tables and/or figures all on one page. If necessary, start the table or figure on a new page, even if this means leaving blank space on the preceding page. If you must split a table over multiple pages, repeat the table headings and continue. It is not necessary to repeat the table title.

Table 4.1: Classroom Tallies

D	A	B	C
E	3	4	7
F	5	8	9

CHAPTER 5: CONCLUSION

Chapter Five, also called "Summary," "Conclusion," or "Recommendations," usually presents a conclusion to the research, offers recommendations to the problem investigated, or discusses implications for future studies.

Bookmarks

A few words about bookmarks. Frontmatter entries, like the Abstract, Acknowledgments and the Table of Contents should appear in the bookmarks – but not in the Table of Contents. The TOC contains only pages that appear after the Table of Contents in the document, usually beginning with the List of Figures. So, bookmark and Table of Contents entries do vary. However, bookmarks should include all major and chapter headings and at least first-level subheadings EXACTLY as they appear in the document (and the TOC). And readers should be able to link to pages within the ETD from all of the bookmarks, the TOC entries, as well as the Lists of Figures and Tables.

APPENDIX A: TITLE OF APPENDIX

• Begin appendix text on the page following the buffer page by using the newpage command.

• Continue Arabic pagination; do not restart page numbering with an appendix

• Use the same style and format for buffer page headings as you do for other body chapter headings.

• Letter, don't number, appendixes (e.g. APPENDIX A, APPENDIX B, etc.)

• If you have only one appendix, do not letter it at all

• Appendixes should follow the margin and other formatting requirements from the rest of the document

APPENDIX B: SECOND APPENDIX

Supplementary documentation.